融合大语言模型与多模态特征的古文命名实体识别[1]

孟佳娜，李丰毅，刘爽，赵迪，王博林
(大连民族大学计算机科学与工程学院 大连 116600)

摘要：［目的/意义］运用命名实体识别技术深入探索古籍文献，推进中文古籍数字化，便于提取和分析重要信息，提升文化遗产的获取与理解，弘扬传统文化。[方法/过程]提出融合大语言模型与多模态特征的古文命名实体识别方法。首先，利用大语言模型进行数据扩充，生成更丰富的样本；然后，使用滑动窗口将文本分割为固定长度的子序列，并将文本子序列输入编码层，得到文本的特征表示；通过卷积神经网络（CNN）提取字形的局部特征，再利用改进的迭代扩张卷积神经网络（IDCNN）提取长距离特征，从而获得字形的全局信息。最后，将文本特征和字形特征在特征感知层进行拼接，形成每个字的综合表示，将拼接后的综合特征传递到CRF层进行序列标注，完成实体预测。以《左传》和CHED_NER为研究语料，构建人名、地名、时间等命名实体识别任务。[结果/结论]实验结果表明，融合大语言模型与多模态特征的古文命名实体识别方法，相比主流的BERT-BiLSTM-CRF方法，F1值分别提升13.32%和1.03%。融合大语言模型与多模态特征的古文命名实体识别方法，能够精准地实现对古籍文本的命名实体识别。

关键词：古文；实体识别；迭代扩张卷积神经网络；大语言模型；特征融合
中图分类号：

1 引言

中华文明传承千年，古籍典籍承载着丰富的历史与文化信息，是探索古代社会、政治、经济与文化的重要资料。然而，随着古籍文献的数量不断增加，如何有效提取这些古文中的关键信息，成为当前古籍数字化和研究的核心难题[1]。特别是在现代信息化和人工智能技术不断发展的背景下，命名实体识别（Named Entity Recognition, NER）作为自然语言处理领域的基础任务之一，为古文的自动化处理与知识挖掘提供了强有力的工具。命名实体识别技术的核心目标是在文本中精准识别并分类诸如人名、地名、时间等具备明确语义的实体。这项技术不仅能够提高古籍的可读性，还能为学术研究提供高效的资料筛选手段[2]。

古籍中的命名实体识别工作具有特别重要的应用价值。一方面，古籍文本蕴含着大量珍贵的历史信息，包括人物关系、地理位置及关键事件的时序安排。这些信息对历史学家、语言学家和文化研究者具有极高的研究价值。通过命名实体识别技术，可以系统地提取这些重要信息，从而加速对古籍内容的自动化处理和深入分析[3]。另一方面，随着现代读者和研究人员对信息快速检索的需求增加，传统的人工处理方式已无法满足这种效率需求。通过自动化的命名实体识别，用户可以迅速从浩如烟海的古籍中找到所需信息，显著提升了古籍的可搜索性与可发现性。

---

[1] 本文系教育部人文社会科学研究规划基金项目“基于知识图谱的中华文化互联网智慧传播研究”(23YJA860010)、辽宁省2024年度社科规划基金项目“政务微博对公共突发事件网络谣言的舆论引导研究”(L24BTQ002)、大连市社科院2024年度调研课题“人工智能视域下政务微博中大众情感风险识别研究”（2024dlsky024）的研究成果之一。
作者简介：孟佳娜，硕士生导师，教授，博士，E-mail:mengjn@dlnu. edu. cn；李丰毅，硕士研究生；刘爽，硕士生导师，教授，博士；赵迪，讲师，博士；王博林，讲师，博士。

尽管命名实体识别技术在现代中文领域已经取得显著进展，然而，应用于古文领域时仍面临一系列独特的挑战。首先，古文与现代汉语在语言结构、词汇使用及语法规则上存在显著差异。例如，古文中的“行”字在不同语境下可能代表“出行”、“施行”或“品行”。这种多义性为模型的语义理解带来了更大的复杂性。此外，古文句式简洁、语法隐含性强，常常省略主语或宾语，使得现代自然语言处理（NLP）模型在捕捉文本的上下文信息与句法边界时面临困难。诸如“知者不惑，仁者不忧，勇者不惧”这样的简洁句式，包含了深刻的思想内容，但模型很难在不丢失语义的情况下准确识别出其中的实体。

另一个挑战在于古文字符和语料的稀缺性。现有的自然语言处理模型大多基于现代汉语语料进行训练，对于古文特有的字符、字形和词性特征，尚未进行充分的特征提取。古籍文本中存在大量的异体字和罕见字符，这些字形特征在语义解读中起着至关重要的作用，但现代 NLP 模型往往难以捕捉这些字形特征[4]。近年来，尽管深度学习和大语言模型的快速发展推动了命名实体识别技术在古文领域的应用，但数据的缺失依然是一个重大挑战。通过引入预训练语言模型（如 BERT、GPT 等），研究人员在一定程度上能够捕捉古文命名实体识别中的复杂上下文信息，提升模型对古文的理解能力。然而，由于古文语料稀缺，训练模型时难以全面覆盖古文特有的字形和语义特征，限制了模型的表现。这一数据缺乏问题使得充分挖掘古文特征和提高识别精度成为当前研究中的一大难点。

针对上述问题挑战，本文以《左传》和 CHED_NER 数据集为实验语料，重点研究古文命名实体识别任务，提出了一种融合字形图像信息及大语言模型的综合方法，构建了包括人名、地名、时间等实体在内的识别模型。通过实验验证，该方法在提升古文命名实体识别的召回率和 F1 值方面表现出显著优势。实验结果证明了融合大语言模型与多模态特征的古文命名实体识别方法的优异性。相较于传统模型，主要贡献如下：

（1）提出了一种融合字形图像特征与大语言模型的多模态古文命名实体识别模型架构（Graph Image Features and LLM-based Named Entity Recognition，GIM-NER）来更好的处理古文命名实体识别任务。

（2）引入了字形特征（图像）增强文本的表达能力，在文本处理中，提供额外的视觉信息，帮助模型更好地理解古文字符的形态变化，从而提高命名实体识别的准确性。

（3）采用回译法，通过荀子大模型[5]进行数据扩充增强，丰富训练数据，提高模型的泛化能力，有助于充分挖掘文本和视觉信息的潜力，提升整体识别能力。

2　相关研究

2.1　基于预训练模型的命名实体识别

近年来，命名实体识别（NER）在深度学习的推动下取得了显著进展，特别是预训练语言模型的应用显著提升了 NER 的性能。在传统模型中，Bi-LSTM-CRF 作为 NER 的基准模型，通过 Bi-LSTM 捕捉上下文语义，再结合 CRF 层建模标签依赖，在多个任务中表现优异。2018 年，Google 团队[6]发布了 BERT（Bidirectional Encoder Representations from Transformers）模型，这标志着 NER 领域的重大突破。BERT 引入了双向 Transformer 结构，能够从文本的双向上下文中学习信息，大幅提升了包括 NER 在内的多项自然语言处理任务的表现。BERT 的成功为预训练语言模型在 NER 任务中的广泛应用奠定了基础。基于 BERT 的成功，Liu 等人[7]提出了 RoBERTa 模型，通过使用更大规模的数据集和更长的预训练时间，进一步提升了 NER 任务的表现。实验表明，RoBERTa 在多项自然语言处理任务中表现优异，尤其是在命名实体识别任务中取得了显著进步。与此同时，研究者们还将 BERT 应用于古文领域。Liu 等人[8]基于《四库全书》繁体语料，提出了 SikuBERT 和 SikuRoBERTa 模型，并专门针对古文命名实体识别任务进行了优化。这些模型在古文数字人文领域中，尤其是史籍实体识别研究中，取得了显著成果。

在古文处理领域，Xu 等人[9]面相《资治通鉴》语料，基于 SikuBERT 预训练模型进行自动摘要实验，结果展示了数字人文技术对古文进行自动摘要任务的可行性和预训练模型对古文进行信息处理的适用性。此外，Wang 等人[10]进一步扩展了 BERT 模型，基于更大规模的古文数据集提出了 Bert-Ancient-Chinese 模型。相比于 SikuBERT 和 SikuRoBERTa 模型，该模型在古文领域的表现更加出色，进一步丰富了古文领域的深度预训练语言模型，并提升了古文命名实体识别等任务的性能。

近年来，研究者们还尝试通过引入字形信息来改进 NER 模型。早期工作如 Liu 等人[11]和 Shao 等人[12]使用 CNN 从汉字图像中提取字形特征。Meng 等人[13]设计了专门的 CNN 结构，用于提取字符的字形特征，并将图像分类作为辅助任务，显著改善模型在中文任务中的表现。Song 和 Sehanobish 以及 Xuan 等人[14]进一步将字形嵌入引入 NER 任务中，结合字形信息显著提升了 NER 模型的性能。相比标准的 BERT 模型，这类方法在处理包含复杂字形信息的中文命名实体识别任务中表现更加优异。

目前，古文命名实体识别在取得显著进展的同时，仍然面临一些挑战。虽然现有的研究大多侧重于基于预训练语言模型的语义特征提取，但对于字形图像特征的有效融合尚未得到充分探索。此外，古文语料稀缺的问题也是限制模型性能提升的重要因素。由于古文数据资源相对匮乏，模型在特定任务中的泛化能力受到限制。针对这些问题，本文的研究重点之一是将字形图像特征与古文命名实体识别任务进行有效融合。通过设计适用于古文的字形特征提取机制，增强模型对字形信息的理解。

2.2 基于数据增强的大语言模型

在大语言模型出现之前，自然语言处理领域的传统数据增强方法（如同义词替换、插入、删除以及回译法）已广泛应用于低资源场景，旨在通过扩展数据量提升模型性能[15]。然而，这些方法虽然简单易行，但在生成数据的多样性和语义一致性上存在不足。Feng 等人[16]和 Hedderich 等人[17]指出，回译法通过跨语言翻译生成多样化样本，但由于依赖翻译模型的质量，往往会引发语义偏差和信息丢失。

近年来，GPT-3 和 GPT-4 等大语言模型的快速发展和应用极大地推进了数据增强技术。LLM 通过大规模数据预训练积累了丰富的语言知识，能够生成高质量、语义一致的文本样本。Chen 等人[18]对比 LLM 和传统方法后指出，LLM 生成的数据不仅准确度更高，且多样性丰富。相比传统的同义词替换和回译法等规则，LLM 可以理解复杂上下文，生成与原始语义一致的多样化变体，尤其在少样本和零样本任务中，LLM 的表现尤为突出，为低资源任务提供了更广泛支持。同时，Zhang 等人[19]进一步探讨了 LLM 如何在生成数据时实现多样性与语义一致性之间的平衡，进一步展示了大模型的灵活性，使其能涵盖更广泛的表达方式。Dai 等人[20]的研究也验证了 ChatGPT 在少样本任务中的数据增强优势，通过生成多样性高且真实的扩展数据，显著提升了文本分类模型的性能。

在数据增强的具体应用中，LLM 结合回译法展现出生成高质量、多样化数据的显著优势。通过跨语言翻译后再回译回原始语言，LLM 能够生成语义一致且多样化的样本。这种方法在保持语义一致性的同时，提供了丰富的表达方式，为低资源任务中的模型训练带来显著帮助。相比传统回译法，LLM 不仅克服了翻译模型的质量局限，还展现出更灵活的语义理解和生成能力，使得回译法在大模型的支持下获得更佳效果。因此，LLM 与回译法的结合在忠实度与多样性之间达到了理想的平衡，为自然语言处理中的大规模数据扩充提供了全新的可能性。本文的研究正是利用大语言模型（如荀子大模型）结合回译法进行数据扩充和增强，以生成更多高质量的古文语料，解决语料稀缺的问题，提升古文命名实体识别模型的性能，为古文处理领域提供更加有效的解决方案。

3 理论与模型

本文提出了一种融合大语言模型与多模态特征的古文命名实体识别模型（GIM-NER）。该模型由数据扩充层、编码层、特征感知层和 CRF[21]层组成。首先，模型通过大语言模型对数据进行扩充，以增强训练数据的多样性和泛化能力，接着，将增强后的数据通过滑动窗口处理分割为固定长度的子序列，以适应模型输入要求。然后，编码层采用预训练的 BERT 模型对文本数据进行编码，生成字级别的上下文表示，再交由双向 LSTM[22]层捕捉序列中的依赖关系。同时，字形图像通过卷积层和迭代扩张卷积神经网络[23]进行处理，提取图像的长距离特征表示。接下来，文本和字形特征被拼接到一起，形成综合特征表示，并传递至 CRF 层进行解码。最后，CRF 层对序列进行全局优化解码，预测出文本中每个字的实体标签，实现命名实体识别。总体模型图如图 1 所示。

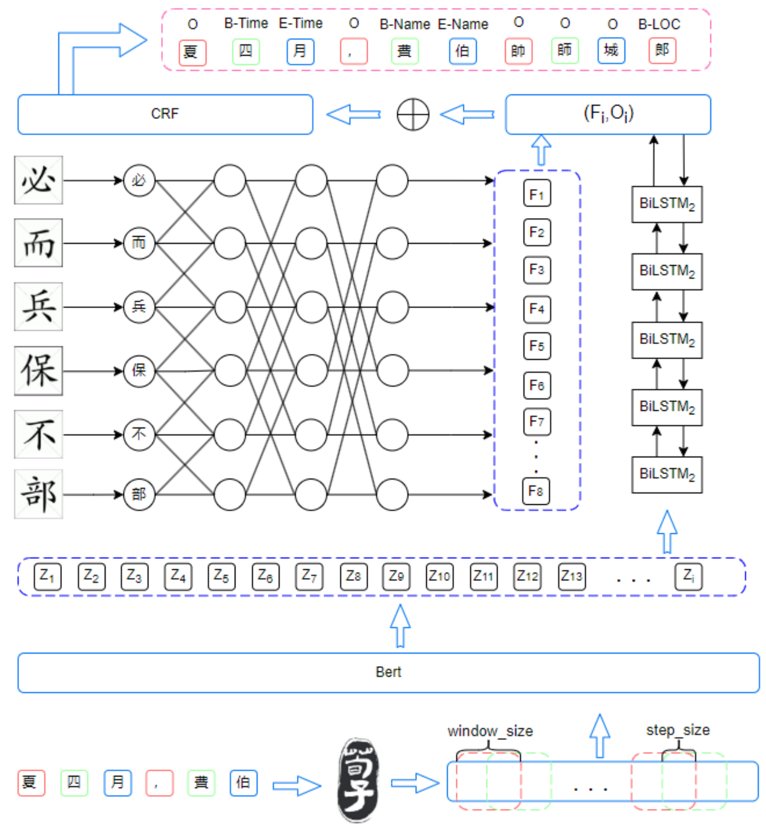


图 1 融合大模型与多模态特征的模型架构

Fig.1 Model architecture that integrates large models and multimodal features

3.1 数据预处理与数据增强

3.1.1 基于大语言模型的数据扩充与增强

在处理古文数据稀缺性问题时，可以有效利用大型语言模型，如 GPT4.0、文心一言等大语言模型，进行数据的扩充和增强。本文的数据集针对古文，借助专门训练的古文领域的模型，荀子大模型，来执行这一任务。本文采用的方法是回译法（round-trip translation）。如图 2 所示。

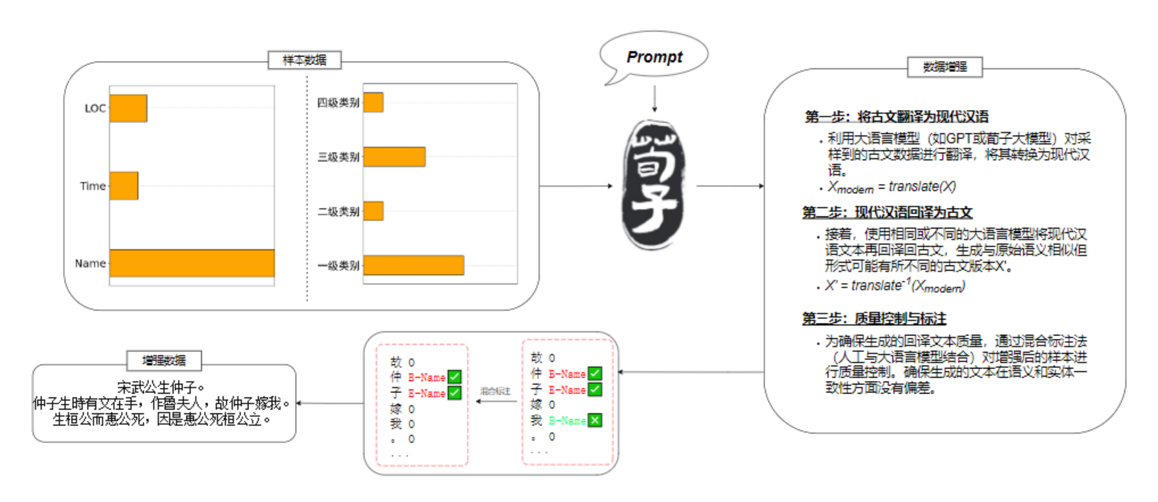


图 1　荀子大语言模型数据增强模型架构

Fig.2 Xunzi Language Model Data Enhancement Model Architecture

图 2 展示了通过回译法生成数据增强的过程。首先将古文文本 x 翻译成现代汉语 xmodern，公式表示如下：

$$x_{\text{modern}} = translate(x) \tag{1}$$

然后将这个现代汉语文本 xmodern 再回译成古文 x′。这一过程不仅帮助生成更多的训练样本，而且可以通过现代汉语的中间层来增强文本的多样性和丰富性。此外，为了保证数据增强的效果和正确性，采用了人工和大模型的共同标注过程。这意味着增强生成的样本不仅仅依赖于自动化模型的输出，还经过人工审核和校正（混合标注）。通过这种人机协作，可以更准确地捕捉到古文的语言特征和文化细节，同时也能有效减少由自动翻译引入的错误。公式表示如下：

$$x' = \text{混合标注}(translate^{-1}(x_{modern})) \tag{2}$$

同时本文还将部分样本数据和增强数据进行实体类型对比如下图。

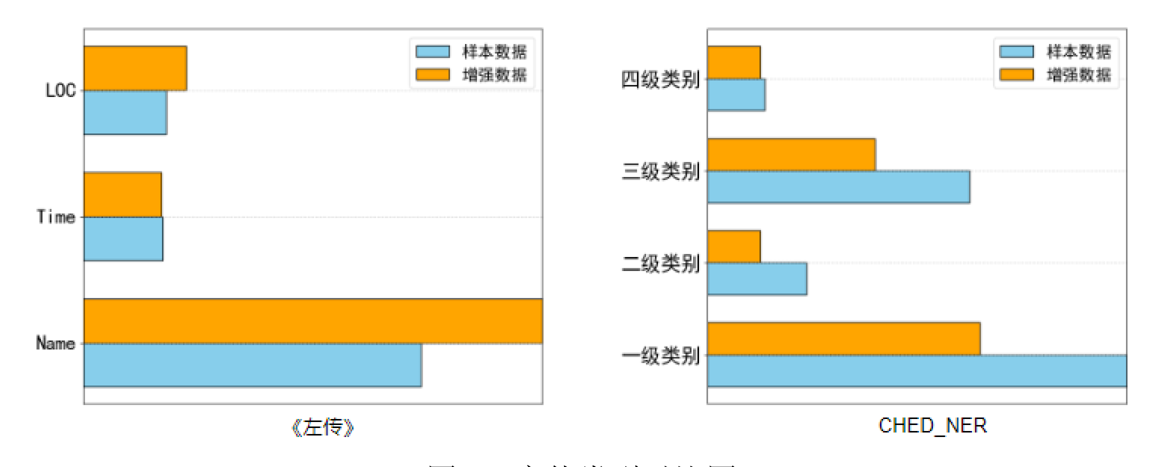


图 3　实体类型对比图

Fig.3 Comparison chart of entity types

图 3 展示了两个数据集中实体类型的对比情况，在《左传》数据集中，增强数据的实体类型普遍多于样本数据，而在 CHED_NER 数据集中，情况正好相反。这个现象可能与大语言模型的生成和样本数据中的类型种类存在一定关联。由于 CHED_NER 数据集中实体类型较多，可能导致大语言模型对语义的把握不如《左传》数据集那样明确。总体来看使用回译法的确可以增强文本的多样性和丰富性。

3.1.2 滑动窗口机制

在处理超长文本输入时，本文采用滑动窗口机制将文本分割为多个固定长度的片段。假设输入文本序列为，$x = [x_1, x_2, x_3 \dots, x_n]$每个片段的长度为 W（窗口大小），步长为 S。通过滑动窗口将超长文本分割为多个片段，每个片段可以表示为：

$$x_{i:i+w} = [x_i, x_{i+1}, x_{i+2}, \dots, x_{i+W-1}] \qquad (3)$$

其中，$x_{i:i+w}$是从位置i开始，长度为 W 的片段。

在这一基础上，将大语言模型生成的扩充数据与原始数据拼接，形成一个新的输入序列。假设扩充后的数据为$x' = [x'_1, x'_2, x'_3, \dots, x'_n]$，将其与原始数据$x$拼接为 X：

$$X_{i:i+w} = [x_1, x_2, x_3, \dots, x_n, x'_1, x'_2, x'_3, \dots, x'_n] \qquad (4)$$

接下来，将对 X 使用滑动窗口，以步长 S 切分成多个长度为 W 的片段：

$$X_{i:i+w} = [X_i, X_{i+1}, X_{i+2}, \dots, X_{i+W-1}] \qquad (5)$$

通过采取这种机制，将使用大语言模型扩充的数据和原始数据结合在一起，利用滑动窗口分割成小片段，为后续的模型处理提供丰富的上下文和信息输入。

3.2 网络层

3.2.1 字形图像特征提取

在处理古文命名实体识别任务时，每个汉字不仅包含语义信息，还包含丰富的字形信息。古文中的汉字往往形态多样，其中包含异体字、罕见字以及一些与现代汉字形态上有显著差异的字符。因此，传统基于语义和上下文的 NER 方法可能难以捕捉到这些复杂的字形特征，导致模型在面对多样化的古文文本时表现不佳。为了解决这一问题，本文设计了一种基于字形特征提取的深度卷积神经网络架构，即带膨胀卷积的迭代扩张卷积网络（IDCNN）。IDCNN 网络可以通过对每个汉字的字形图像进行卷积操作，有效提取其中蕴含的结构信息和细节特征，从而增强模型对古文字形的理解能力。这种方法尤其适用于古文文本，因为它能够扩展感受视野，确保在提取特征时保留汉字的细微差别。通过字形图像特征，IDCNN 为古文 NER 任务提供了更加精准和鲁棒的解决方案。

假设输入的字形图像为l_j，通过 IDCNN 提取的特征表示为F_j。这个过程可以通过以下公式表示：对于第j个汉字的字形图像l_j，卷积核$W^{(l)}$和膨胀率r_l的卷积操作为：

$$F_j^{(L)} = \sigma(w^{(l)} * r_l F_j^{(l-1)} + b^{(l)}) \qquad (6)$$

其中$* r_l$表示带有膨胀率r_l的卷积操作，这种膨胀率有助于扩大卷积视野，以捕捉更丰富的字形图像细节。

通过多层卷积网络提取的字形图像特征表示为：

$$F_j = IDCNN(i_j) \in R^{C \times H \times W} \qquad (7)$$

公式（7）表示的是将输入的字形图像l_j经过 IDCNN 网络后得到的输出特征图。F_j表示第j个汉字的字形图像特征，是一个三维张量，用于表示汉字在特征空间中的表示。$IDCNN(i_j)$表示 IDCNN 模型对输入的字形图像l_j进行卷积操作后的输出结果。IDCNN 通过多层卷积逐层提取汉字的边缘、笔画和空间结构等信息，从而将这些细节特征编码到特征图中。其中输出特征图的维度说明：C 是输出的通道数（channels），表示 IDCNN 提取到的特征类型数量。通道数越多，模型捕捉到的信息越丰富，每个通道可视为不同的特征检测器，如笔画方向或汉字结构。H 和 W 分别为特征图的高度和宽度，这些值由输入图像大小、卷积核大小、步长

及膨胀率决定。IDCNN 通过膨胀卷积扩大感受野，保留更多空间信息，以捕捉汉字的整体结构和细节特征。

3.2.2 文本特征提取与融合

在处理现代文本的命名实体识别（NER）任务中，预训练的 BERT 模型和双向长短时记忆网络（BiLSTM）的结合提供了一种强大的方法来捕获每个 token 的深层上下文信息。预训练的 BERT 模型首先对文本进行编码，为每个 token 生成丰富的嵌入向量z_i这些向量能够反映词汇之间的复杂语义关系和各种语境变化。公式如下：

$$Z = BERT(X) \tag{8}$$

BiLSTM 进一步增强这些特征，通过从两个方向分别处理文本信息，生成更全面的双向上下文特征。前向 LSTM 捕捉从左到右的语境依赖，其计算公式为：

$$\vec{h}_i = \sigma\left(W_z Z_i + W_h \vec{h}_{i-1} + b_h\right) \tag{9}$$

而后向 LSTM 则从右到左对文本进行处理，提供对未来词汇的信息，公式为:

$$\overleftarrow{h}_i = \sigma\left(W_z Z_i + W_h \overleftarrow{h}_{i+1} + b_h\right) \tag{10}$$

BiLSTM 的最终输出通过前向和后向隐状态的拼接得到:

$$O_i = \vec{h}_i \oplus \overleftarrow{h}_i \tag{11}$$

为每个词提供了其在上下文中的完整语义表征。为了进一步提高模型的表现，特别是在处理形态多样的古文或具有丰富字形变体的语料时，文本的上下文特征与字形图像特征进行融合。每个字符的视觉特征F_i通过深度卷积网络处理，提取关键的形态信息，如笔画、结构和样式。这些字形图像特征与 BiLSTM 的输出O_i在特征层面进行拼接，形成融合后的特征表示：

$$Z_i' = O_i \oplus F_i \tag{12}$$

既包含了丰富的语义信息，也融入了字形的独特视觉信息，从而使模型能够在处理如异体字或罕见字符时展现更高的准确性和鲁棒性。通过这种综合特征的融合，模型在面对复杂文本环境时能提供更为精确的实体识别和文本分析结果。

3.2.3 序列标注、解码与损失计算

在序列标注任务中，模型通过结合文本和视觉特征，提高对实体的识别能力。具体来说是文本特征与进一步加工的字形图像特征进行融合，形成综合特征Z_i'。这些融合后的特征包含了文本的上下文信息和视觉信息，提供了丰富的数据输入给后续层。

全连接层W_{fc}接收这些融合特征，并将其映射到不同的标签类别维度，为每个 token 生成一个未归一化的标签得分（logits）：

$$P_i = W_{fc} Z_i' + b_{fc} \tag{13}$$

这里的P_i是第 i 个 token 的得分，这些得分随后用于条件随机场（CRF）层的序列解码。

在序列标注任务中，模型使用 CRF 层来解码标签序列，优化整体的标签选择过程。CRF 层不仅利用每个 token 的得分，还考虑标签间的转移概率，确保输出序列的全局一致性。通过最小化负对数似然，模型计算得到的损失函数主要基于真实标签序列y_{true}的概率。此外，如果存在额外的损失项，如边界检测或特征识别，这些将与主损失函数通过超参数$\lambda_1, \lambda_2, \lambda_3$加权合并，以支持多任务学习，从而使模型能够在处理复杂文本时表现出更高的准确性和适应性。

4 实验分析

4.1 实验数据集

本文一共采用两个数据集《左传》还有 CHED_NER[24]数据集，其中左传是第一届古汉语分词与词性标注评测 EvaHan2022 所使用的数据集，其中训练集一共包含 13400 条句子，验证集 1600 条句子，测试集 1600 条句子，训练集中的实体类型主要有三种：人名（Name）、地名（Loc）、时间（Time）。CHED_NER 数据集包含来自《二十四史》中 61 篇文档，61 个历史人物。其中训练集一共包含 5600 条句子，验证集 1200 条句子，测试集 1200 条句子。此

数据集古文历史事件类型之间是有层次关系的，分为一二三四级类别，例如："命"在某些地方是四级类别 B-交流-个人交流-诏令-命令，另外事件大类之间也是具有联系的，大体上包括 9 大类和 67 小类的古文历史事件类型。表 1 具体展示了《左传》和 CHED_NER 数据集语句规模统计的详细情况，表 2 和表 3 分别展示了《左传》和 CHED_NER 数据集实体分布情况。

表 1　《左传》& CHED_NER 数据集统计表

Table1 Data Set Statistical Table of Zuo Zhuan And CHED_NER

数据集	类型	训练集	验证集	测试集
《左传》	Sentence	13.4K	1.6K	1.6K
	Char	194.9K	28.2K	34.7K
《CHED_NER》	Sentence	5.6K	1.2K	1.2K
	Char	123.1K	26.0K	26.1K

表 2　《左传》实体分布统计表

Table2 Entity Distribution Statistical Table of Zuo Zhuan

实体种类	训练集	验证集	测试集
人名	10662	2107	2157
地名	5199	1486	1875
时间	1474	472	459

表 3　CHED_NER 实体分布统计表

Table3 Entity Distribution Statistical Table of CHED_NER

实体种类	训练集	验证集	测试集
一级类别	10632	2200	2277
二级类别	3010	604	630
三级类别	6123	1304	1327
四级类别	1499	292	320

《左传》数据集使用 BIOES 标注体系进行序列标注。在 BIOES 序列标注体系中，B 代表实体的起始位置，I 代表实体的中间位置，O 代表非实体部分，E 代表实体终止位置，S 代表单独字为一个实体。CHED_NER 数据集是古汉语历史事件检测数据集 (CHED) 转换格式后的子集，将事件检测任务建模为序列标注任务，专门用于命名实体识别 (NER) 任务。该数据集采用 BIO（Begin, Inside, Outside）格式，其中每个词被标注为事件触发词（B-标签）、触发词的一部分（I-标签），或者非事件相关的词（O）。序列标注情况如表 4、5 所示。

表 4　《左传》序列标注方式

Table4 Sequence labeling of Zuo Zhuan

字	含义	序列标签
夏	非实体词	O
四	时间	E-Time
月	时间	E-Time
，	非实体词	O
费	人名	B-Name
伯	人名	E-Name
帥	非实体词	O
師	非实体词	O
城	非实体词	O
郎	地名	B-LOC
。	非实体词	O

表 5　CHED_NER 序列标注方式

Table5 Sequence labeling of CHED_NER

字	含义	序列标签
甲	非实体词	0
辰	非实体词	0
，	非实体词	0
命	四级类别	B-交流-个人交流-诏令-命令
诸	非实体词	0
道	非实体词	0
进	三级类别	B-军事-备战-出兵
讨	四级类别	B-军事-作战-攻击-征伐
。	非实体词	0

4.2　评价指标

在自然语言处理任务中，常用的评价指标包括准确率（Accuracy）、精确率（Precision）、召回率（Recall）和综合评价指标（F1-Score）。这些指标用于衡量模型在处理任务时的表现，帮助评估分类结果的质量。

（1）准确率（Accuracy）

准确率表示模型预测正确的样本占总样本的比例。它衡量了模型整体上的正确率，计算公式如下：

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \tag{14}$$

其中，TP 为真正例，TN 为真负例，FP 为假正例，FN 为假负例。准确率能够反映模型总体的预测效果，但在类别不平衡问题中可能不足以体现模型的分类能力。

（2）精确率（Precision）

精确率表示在所有被预测为正例的样本中，实际为正例的比例。它衡量了模型对正例预测的准确性，计算公式如下：

$$Precision = \frac{TP}{TP+FP} \tag{15}$$

精确率适合关注假正例较多的场景，能够反映模型的预测精度。

（3）召回率（Recall）

召回率表示在所有实际为正例的样本中，模型正确预测为正例的比例。它反映了模型对正例的覆盖能力，计算公式如下：

$$Recall = \frac{TP}{TP+FN} \tag{16}$$

召回率关注的是模型识别正例的能力，适用于需要减少漏检的场景。

（4）综合评价指标（F1-Score）

为了兼顾精确率和召回率，常常使用 F1-Score 作为综合评价指标。F1-Score 是精确率和召回率的调和平均值，计算公式如下：

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{17}$$

F1-Score 提供了精确率与召回率之间的平衡，适用于需要均衡考虑两者的场景。

4.3 实验环境与实验参数设置

实验环境为 Window10 操作系统、12 vCPU Intel(R) Xeon(R) Platinum 8255C CPU @ 2.50GHz、16 GB RAM 以及 NVIDIA GeForce GTX 3080 SUPER 图形处理器。模型框架为 Python3.8、torch1.10。为了模型性能达到最佳的效果，本文经过多次训练，对模型中需要手动输入的参数值进行调整确定其他参数详见表 6。

表 6 参数设置表

Table6 Parameter Settings Table

超参数	值
学习率	1e-5
Batch 大小	16
迭代次数	50
卷积核大小	3×3
卷积膨胀率	[1,2,4]
BiLSTM 隐藏层维度	128
输入句子最大长度	100
BERT 隐藏层维度	768
图像特征维度	32×32
梯度下降优化器	Adam
Dropout	0.5
分词工具	HanLP

4.4 实验结果与分析

4.4.1 基线模型对比实验

以《左传》和 CHED_NER 数据集为研究语料，选取 BERT-CRF、BiLSTM-CRF、IDCNN-CRF、BERT-IDCNN-CRF、BERT-BiLSTM-CRF 五种基线模型进行实验对比，实验结果如表 7 所示：

表 7 基线模型实验对比

Table7 Baseline model experiment comparison

对比基线模型	《左传》			CHED_NER		
	P（%）	R（%）	F1（%）	P（%）	R（%）	F1（%）
BERT-CRF	78.50	72.84	75.14	58.40	51.35	53.15
BiLSTM-CRF	58.50	50.25	53.74	60.77	52.91	54.76
IDCNN-CRF	68.84	49.07	56.44	55.50	46.88	49.44
BERT-IDCNN-CRF	84.48	72.47	77.52	65.56	73.48	69.12
BERT-BiLSTM-CRF	**89.17**	70.12	**78.72**	**72.93**	**77.68**	**75.23**

由表 7 可知，从实验结果可以看出，BERT 作为预训练模型在古文命名实体识别任务中具有显著的优势。无论是与 CRF 直接结合，还是与 IDCNN 或 BiLSTM 相结合，BERT 都展现了强大的性能提升，尤其在与 BiLSTM 和 CRF 结合时，F1 分数达到了 78.72 和 75.23，整体表现最好。而相比之下，IDCNN 虽然在特征提取速度上优于 BiLSTM，但在复杂句法结构的处理上略显不足，尤其是在较为复杂的上下文捕捉能力上不如 BiLSTM。因此，BiLSTM 更适合作为特征提取的模块。综合来看，BERT 与 BiLSTM 和 CRF 组合的模型在处理上下文依赖性强的任务中表现最为出色，是目前最佳的基线模型组合。

4.4.2 实验对比

在《左传》和 CHED_NER 数据集上，本文提出的 GIM-NER 模型与已有的多种模型进行对比，实验结果如表 8 所示：

（1）TENER[25]：Yan 等人提出的 TENER 模型通过引入方向和距离的相对位置编码及非缩放的注意力机制，提升了 Transformer 在命名实体识别任务中的表现。TENER 结合字符级和词级特征，能够有效捕捉长程依赖，在多个中英文数据集上超越了传统的 BiLSTM 模型。

（2）FLAT[26]：Li 等人通过将汉字与词语的信息整合为扁平的结构，并采用 Transformer 的自注意力机制，在保持并处理词边界和语义信息的同时，显著提升了中文命名实体识别任务的性能和计算效率。

（3）SIMP[27]：Ma 等人通过将词典信息整合到字符表示层中，简化了中文命名实体识别中词典使用的复杂性，提升了推理速度和模型性能。相比复杂的 Lattice-LSTM 架构，SoftLexicon 通过在字符表示中直接引入词典匹配信息，大幅加快了处理速度，并可以轻松应用于不同的神经网络架构，如 BiLSTM、CNN 和 Transformer，同时保持了较好的识别效果。

（4）MECT[28]：Wu 等人通过引入基于多元嵌入的交叉 Transformer 网络，将汉字的部首结构信息融入命名实体识别任务中，增强了模型对汉字语义和边界信息的捕捉能力，并在多个基准数据集上表现出优越性。

（5）HGN[29]：Hu 等人通过结合 Transformer 的全球信息提取能力和多窗口递归模块的局部信息捕捉能力，提升了命名实体识别的性能。使用 Transformer 编码器获取全局上下文信息，并通过多窗口滑动机制提取局部特征及相对位置信息，最后通过多窗口注意力机制将这些信息融合，显著提高了多个通用和生物医学领域数据集上的 NER 表现。

（6）MarkBERT[30]：Li 等人提出了一种通过在字符之间插入标记的方式来增强 BERT 对汉字词界限信息的理解。不同于传统基于字符或词的模型，MarkBERT 利用特殊的标记插入策略，无论是高频词还是低频词，均能通过标记来提供词的边界信息，从而避免了词汇表覆盖不全的问题。

表 8　不同模型对比实验

Table8 Comparative Experiment of Different Model

模型	《左传》			CHED_NER		
	P（%）	R（%）	F1（%）	P（%）	R（%）	F1（%）
TENER	71.70	59.66	65.13	67.28	75.10	70.98
FLAT[25]	88.69	75.83	81.76	73.60	76.82	75.18
SIMP[26]	89.73	75.96	82.13	73.44	69.67	72.47
MECT[27]	**90.07**	75.04	81.85	72.97	62.75	66.89
HGN[28]	88.53	79.12	83.56	72.14	72.07	70.55
MarkBERT[29]	82.18	77.70	79.88	**74.47**	77.72	76.07
GIM-NER	87.96	**83.44**	**85.35**	74.36	**78.62**	**76.26**

由表 8 可知，本文提出的 GIM-NER 模型在《左传》和 CHED_NER 数据集上表现优异，F1 值达到了 85.35%，召回率 83.44%。相比之下，其他模型虽然各有所长，但仍有局限：FLAT 模型提升了词汇信息质量，但未充分利用字形和语义特征；SIMP 模型增强了字词特征表示，但缺乏对汉字结构的深度结合；MECT 模型经过引入了汉字结构信息，准确率达到了 90.07%，这样的结构信息可能使得模型更易于捕捉汉字语义的细微差别，从而提升了识别的准确性，但对潜在特征的挖掘不够充分；HGN 模型结合了局部和全局特征，但在全面捕捉语义上存在局限；MarkBERT 模型通过插入标记解决了词界限问题，但无法充分挖掘复杂字形特征。相

比之下，GIM-NER 模型结合了字形特征与大模型数据增强，全面捕捉了字形、结构和上下文特征，并泛化了模型，使得其性能优于其他模型。

4.4.3 消融实验

为进一步验证 GIM-NER 模型各个模块的有效性，在《左传》和 CHED_NER 数据集进行不同子模块消融实验。实验结果如表 9 所示。

（1）GIM-NER：加入字形图像特征、使用大模型扩充增强

（2）-字形图像:去除字形图像特征

（3）-LLM：不使用大模型进行数据扩充增强。

（4）-字形图像-LLM：去除字形图特征，不使用大模型扩充的数据进行增强。

表 9　消融实验

Table9 Ablation experiment

模型	《左传》			CHED_NER		
	P（%）	R（%）	F1（%）	P（%）	R（%）	F1（%）
GIM-NER	87.96	83.44	85.35	74.36	78.62	76.26
-字形图像	84.31	79.60	82.32	73.28	77.33	75.05
-LLM	87.44	82.89	84.83	73.47	77.45	75.22
-字形图像-LLM	89.17	70.12	78.72	74.73	75.42	74.87

由表 9 可以看出，去掉字形图像时，《左传》数据集的 F1 值下降了 3.03%，而 CHED_NER 数据集下降了 1.21%，这表明字形图像在《左传》数据集上起到了更显著的作用。由于《左传》数据集包含丰富的汉字形态特征，字形图像的加入能更好地帮助模型识别复杂的文字结构，因此去除字形图像对其影响较大。而在 CHED_NER 数据集中，实体种类较为多样，字形特征对模型的帮助相对有限，所以去除后影响较小。另一方面，去掉大语言模型扩充增强时，《左传》数据集的 F1 值下降了 0.52%，CHED_NER 数据集的 F1 下降了 1.04%。大语言模型扩充为数据集提供了更多上下文语境，尤其对多类型实体的识别更有帮助，因此在 CHED_NER 数据集上影响更为显著。《左传》数据集由于文本结构相对固定，对扩充增强的依赖性较低，因此表现出较小的下降幅度。

去掉字形图像和大语言模型两者后，模型的准确率略有提高，原因可能在于特征简化确实有助于减少过拟合和噪声干扰，导致准确率略有提高，但这并不代表整体性能更好。本文的 GIM-NER 模型在保持字形图像和 LLM 特征后，显著提高了召回率和 F1 值，表明它在复杂文本和字符结构中具有更强的特征捕捉能力。F1 值是衡量模型综合表现的核心指标，而准确率的提升只是特征简化后的一个附带效果。GIM-NER 模型通过字形与上下文特征的互补，在识别准确性和全面性上都优于去除特征后的模型，充分展示了其在复杂文本处理中的优势。

总体而言，实验结果验证了字形图像和大语言模型扩充增强对模型性能的提升效果，特别是对包含复杂汉字结构的文本和具有特定语境需求的数据集，方法的有效性尤为突出。

5　结语

本文提出了一种适用于古文命名实体识别的 GIM-NER 模型，该模型包含数据扩充层、编码层、特征感知层和 CRF 层。通过利用大语言模型进行数据扩充，模型增强了数据的多样性和泛化能力。此外，采用滑动窗口技术处理增强后的数据，以满足模型输入要求。滑动窗口方法不仅能够处理长文本，还能在确保上下文信息完整性的同时，提高数据的多样性。这一技术对古文命名实体识别尤为重要，因为古文中的许多句式简洁，省略了主语或宾语，传统的长文本处理方法往往难以准确捕捉其上下文关系。在编码层中，模型通过预训练的 BERT 对文本进行字符级别的上下文编码，并引入双向 LSTM 层以捕捉序列依赖关系。同时，字形图像通过卷积层和改进的 IDCNN 提取有效的长距离特征。这不仅弥补了纯文本特征的不足，

也使得模型能够在面对古文特有的书写和字符变动时，保持较高的准确度和稳定性。文本特征和图像特征的拼接，使得模型在 CRF 层进行有效的序列解码，从而以全局优化的方式准确预测文本中的实体标签。GIM-NER 模型在《左传》和 CHED_NER 两个数据集上均表现出色，尤其是在召回率和 F1 分数方面，充分验证了其在古文命名实体识别任务中的有效性与先进性。

未来，我们计划在模型优化上进一步引入先进的图像特征提取方法和多模态特征融合技术，以增强模型对复杂字形结构和文本语义的理解，从而实现在低资源环境下实现更高的识别精度与稳定性，为古文命名实体识别和数字化应用奠定更扎实的基础。

参考文献

[1] 李明杰,张纤柯,陈梦石.古籍数字化研究进展述评(2009-2019)[J].图书情报工作,2020,64(06):130-137.DOI:10.13266/j.issn.0252-3116.2020.06.015.

[2] Su W, Zhao D, Meng J, et al. Named Entity Recognition for Ancient Chinese Based on Knowledge Embedding[C]//2023 3rd International Conference on Digital Society and Intelligent Systems (DSInS). IEEE, 2023: 429-432.

[3] 鞠孜涵,白如江,张玉洁,等.数字人文视域下古籍数据库建设关键技术研究——兼评稷下学文献资料数据库的建设思路[J].图书情报工作,2022,66(19):4-14.DOI:10.13266/j.issn.0252-3116.2022.19.001.

[4] 孟伟伦,郭景峰,邢珂萱,等.基于字形特征的中文医学命名实体识别方法[J].电子学报,2024,52(06):1945-1954.

[5] 光明日报.古籍智能化工具“荀子古籍大语言模型”在京发布[EB/OL]. [2023-12-06]. https://app.gmdaily.cn/as/opened/n/20b439f2c8634d408e1ac1db85481139.

[6] Devlin J. Bert: Pre-training of deep bidirectional transformers for language understanding[J]. arXiv preprint arXiv:1810.04805, 2018.

[7] Liu Y. Roberta: A robustly optimized bert pretraining approach[J]. arXiv preprint arXiv:1907.11692, 2019, 364.

[8] 刘江峰,冯钰童,王东波等. 数字人文视域下 SikuBERT 增强的史籍实体识别研究[J].图书馆论坛,2022,42(10):61-72.

[9] 徐润华,王东波,刘欢,等.面向古籍数字人文的《资治通鉴》自动摘要研究——以 SikuBERT 预训练模型为例[J].图书馆论坛,2022,42(12):129-137.

[10] Wang P, Ren Z. The uncertainty-based retrieval framework for Ancient Chinese CWS and POS[J]. arxiv preprint arxiv:2310.08496, 2023.

[11] Liu F, Lu H, Lo C, et al. Learning character-level compositionality with visual features[J]. arXiv preprint arXiv:1704.04859, 2017.

[12] Shao Y, Hardmeier C, Tiedemann J, et al. Character-based joint segmentation and POS tagging for Chinese using bidirectional RNN-CRF[J]. arXiv preprint arXiv:1704.01314, 2017.

[13] Meng Y, Wu W, Wang F, et al. Glyce: Glyph-vectors for chinese character representations[J]. Advances in Neural Information Processing Systems, 2019, 32.

[14] Xuan Z, Bao R, Jiang S. FGN: Fusion glyph network for Chinese named entity recognition[C]//Knowledge Graph and Semantic Computing: Knowledge Graph and Cognitive Intelligence: 5th China Conference, CCKS 2020, Nanchang, China, November 12–15, 2020, Revised Selected Papers. Springer Singapore, 2021: 28-40.

[15] 许钦亚,薛秋红,钱力,等.融合ChatGPT数据增强的学术论文语步识别方法研究[J].图书情报工作,2024,68(17):84-94.DOI:10.13266/j.issn.0252-3116.2024.17.007.

[16] Feng S Y, Gangal V, Wei J, et al. A survey of data augmentation approaches for NLP[J]. arXiv preprint arXiv:2105.03075, 2021.
[17] Hedderich M A, Lange L, Adel H, et al. A survey on recent approaches for natural language processing in low-resource scenarios[J]. arXiv preprint arXiv:2010.12309, 2020.
[18] Chen J, Tam D, Raffel C, et al. An empirical survey of data augmentation for limited data learning in nlp[J]. Transactions of the Association for Computational Linguistics, 2023, 11: 191-211.
[19] Zhang R, Wang Y S, Yang Y. Generation-driven Contrastive Self-training for Zero-shot Text Classification with Instruction-following LLM[J]. arXiv preprint arXiv:2304.11872, 2023.
[20] Dai H, Liu Z, Liao W, et al. Auggpt: Leveraging chatgpt for text data augmentation[J]. arXiv preprint arXiv:2302.13007, 2023.
[21] 杜悦,王东波,江川,等.数字人文下的典籍深度学习实体自动识别模型构建及应用研究[J].图书情报工作,2021,65(03):100-108.DOI:10.13266/j.issn.0252-3116.2021.03.013.
[22] 余传明.基于深度循环神经网络的跨领域文本情感分析[J].图书情报工作,2018,62(11):23-34.DOI:10.13266/j.issn.0252-3116.2018.11.003.
[23] 韩普,顾亮.基于混合深度学习的中文医学实体抽取研究[J].图书情报工作,2022,66(14):119-127.DOI:10.13266/j.issn.0252-3116.2022.14.012.
[24] Congcong W, Zhenbing F, Shutan H, et al. CHED: A Cross-Historical Dataset with a Logical Event Schema for Classical Chinese Event Detection[C]//Proceedings of the 22nd Chinese National Conference on Computational Linguistics. 2023: 875-888.
[25] Yan H, Deng B, Li X, et al. TENER: adapting transformer encoder for named entity recognition[J]. arXiv preprint arXiv:1911.04474, 2019.
[26] Li X, Yan H, Qiu X, et al. FLAT: Chinese NER using flat-lattice transformer[J]. arXiv preprint arXiv:2004.11795, 2020.
[27] Ma R, Peng M, Zhang Q, et al. Simplify the usage of lexicon in Chinese NER[J]. arXiv preprint arXiv:1908.05969, 2019.
[28] Wu S, Song X, Feng Z. MECT: Multi-metadata embedding based cross-transformer for Chinese named entity recognition[J]. arXiv preprint arXiv:2107.05418, 2021.
[29] Hu J, Shen Y, Liu Y, et al. Hero-gang neural model for named entity recognition[J]. arXiv preprint arXiv:2205.07177, 2022.
[30] Li L, Dai Y, Tang D, et al. Markbert: Marking word boundaries improves chinese bert[C]//CCF International Conference on Natural Language Processing and Chinese Computing. Cham: Springer Nature Switzerland, 2023: 325-336.

Named Entity Recognition for Ancient Chinese Texts Using LLMs and Multimodal Features

MENG Jiana，LI Fengyi，LIU Shuang，ZHAO Di，WANG Bolin

（School of Computer Science and Enginnering，Danlian Minzu University，Dalian 116600，China）

Abstract[Purpose/Significance]This study aims to explore ancient Chinese texts using Named Entity Recognition (NER) technology, promote the digitization of ancient Chinese texts, facilitate the extraction and analysis of important information, enhance the acquisition and understanding of cultural heritage, and promote traditional culture. [Method/Process]We propose a method for NER in ancient Chinese texts that integrates large language models with multimodal features. First, we utilize a large language model for data augmentation to generate richer samples. Then, we segment the text

into fixed-length subsequences using a sliding window approach and input these subsequences into an encoding layer to obtain feature representations of the text. Convolutional Neural Networks (CNN) are employed to extract local features of the character shapes, and an improved Iterative Dilated Convolutional Neural Network (IDCNN) is used to capture long-range features, thereby obtaining global information of the character shapes. Finally, the text features and shape features are concatenated at a feature perception layer to form a comprehensive representation for each character, and the concatenated comprehensive features are passed to a CRF layer for sequence labeling to complete entity prediction. Using "Zuo Zhuan" and CHED_NER as the research corpus, we constructed tasks for identifying named entities such as personal names, geographical names, and temporal expressions. [Result/Conclusion]Experimental results show that the ancient Chinese text named entity recognition method that integrates large language models and multimodal features has improved F1 values by 13.32% and 1.03% respectively compared to the mainstream BERT-BiLSTM-CRF method.The proposed method for NER in ancient Chinese texts, which integrates large language models with multimodal features, can accurately achieve named entity recognition in ancient Chinese texts.
Keywords: Ancient Chinese Texts； Entity Recognition； Iterative Dilated Convolutional Neural Network； Large Language Model； Feature Fusion